ISSN 0868–5886 *НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2014, том 24, № 3, с. 87–91*

МАТЕМАТИЧЕСКИЕ АНАЛИЗ И МОДЕЛИРОВАНИЕ В ПРИБОРОСТРОЕНИИ

УДК 543.426; 543.9

© Д. А. Белов, Ю. В. Белов, В. В. Манойлов, В. Е. Курочкин

СПОСОБЫ ОБРАБОТКИ РЕЗУЛЬТАТОВ ГЕНЕТИЧЕСКИХ АНАЛИЗОВ

Выполнено моделирование генетических сигналов для оценки эффективности способов обработки информации. Рассмотрены возможности различных способов коррекции базовой линии, фильтрации и повышения разрешающей способности при генетических анализах.

Кл. сл.: ДНК, генетический анализатор, флуоресцентная детекция

ВВЕДЕНИЕ

Определение нуклеотидной последовательности (секвенирование ДНК) и фрагментный анализ являются основными задачами, решаемыми с помощью генетического анализатора, основанного на принципе капиллярного электрофореза [1].

Для раздельного детектирования фрагментов ДНК на конец каждого фрагмента помещается соответствующая флуоресцентная метка. При секвенировании на выходах четырех цветовых каналов флуоресцентного детектора регистрируются в цифровом виде сигналы зависимости интенсивности флуоресценции от времени, которые графически регистрируются в виде последовательностей пиков, соответствующих нуклеотидам A, C, G, T. Задача определения нуклеотидной последовательности решается путем измерения положения во времени пиков в каждом канале флуоресцентного детектора, присвоения им буквенных обозначений и суммирования результатов в виде буквенной последовательности.

Для фрагментного анализа ДНК флуоресцентный детектор имеет пятый цветовой канал, на выходе которого могут быть получены сигналы калибровочной смеси фрагментов ДНК известной длины [2].

В настоящей статье выполнено моделирование сигналов детектора. Модельные сигналы использованы для сравнения эффективности способов аппроксимации базовой линии, фильтрации сигналов, повышения разрешающей способности и определения параметров пиков при анализах нуклеиновых кислот.

МОДЕЛИРОВАНИЕ СИГНАЛОВ ДЕТЕКТОРА

Имитационное моделирование сигналов фрагментов ДНК одного из каналов флуоресцентного детектора выполнено в виде последовательности пиков, в которой некоторые пики произвольно располагаются поодиночке, а другие — группами, включающими 2, 3 и 4 соседних пика. Форма пиков определяется функцией Гаусса, график модельных сигналов имеет 3 участка по 1000 временны́х точек, эти участки отличаются шириной пиков.

В терминах программы Excel функция выбранной последовательности пиков вычислена в виде электронной таблицы в столбце `B`. Ячейка `B1` имеет вид:

`B1=M*2.72^(-((X1-A1)^2)/(2*σ^2))+`
`+2.72^(-((X2-A1)^2)/(2*σ^2))+`
`+2.72^(-((X3-A1)^2)/(2*σ^2))+` и т. д.,

где `X1=60, X2=180, …, X39=2960` — положения центров 39 пиков;
`A1=1, A2=2, …, A3000=3000` — порядковый номер ячейки в столбце `A`;
`B1, B2, …` — относительная интенсивность сигнала флуоресценции (в ячейках столбца `B`), все рассчитываются аналогично `B1`;
`M` = 1000 о.е. — высота пиков (в относительных единицах флуоресценции);
2.72 — приближенное значение числа *e*;
`σ^2` — степень 2 числа σ;
σ = 6, 8 и 9 на 1, 2 и 3 участках графика модельных сигналов;
минимальное расстояние между пиками — 20 точек.

Дрейф базовой линии моделировался в столбце `C` полиномом второй степени:

`C= m*A+n*A^2,`

где `m` и `n` — коэффициенты, выбранные из условий `m×A3000= 500` и `n×A3000^2=1000` (суммарный дрейф в конце графика равен 1500 о.е.).

В качестве модели шумового сигнала использовалась последовательность случайных чисел

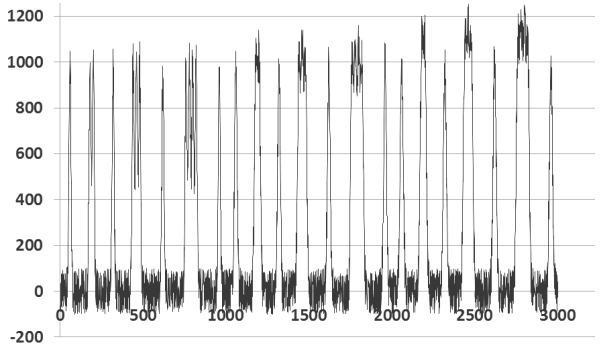

Рис. 1. Модельный сигнал.
Горизонтальная ось — номер отсчета, вертикальная ось — интенсивность флуоресценции в относительных единицах (о.е.)

с максимальной интенсивностью ± 100 о.е.

Модельный сигнал, объединяющий пики и шум, приведен на рис. 1. Модельные сигналы пиков, базовой линии, шумового сигнала и их суммы сохранялись в текстовом редакторе и использовались далее в программе `MATLAB`.

Далее анализируются методы предварительной обработки сигналов.

КОРРЕКЦИЯ БАЗОВОЙ ЛИНИИ

В 2013 г. Институтом аналитического приборостроения РАН (Санкт-Петербург) совместно с ЗАО "Синтол" (Москва) был разработан и успешно испытан генетический анализатор НАНОФОР-05 [3].

Обычно обработка сигналов детектора начинается с устранения начального смещения и дрейфа базовой линии отдельно в каждом цветовом канале. В генетическом анализаторе НАНОФОР-05 был применен традиционный способ вычисления базовой линии — медианная фильтрация. В последовательности ДНК большое количество пиков находится на близком расстоянии друг к другу, поэтому возникают большие погрешности при таком способе коррекции базовой линии и искажаются сигналы ДНК. В статье [4] предложен способ вычисления и коррекции базовой линии, обладающий значительно меньшими погрешностями, однако при этом не учитывалось влияние шумов детектора. Этот способ был реализован в среде Excel, поэтому возникли трудности при его применении в генетическом анализаторе.

Для вычисления и коррекции базовой линии предлагается усовершенствованный автоматический способ, реализованный в программе `MATLAB`.

Обработка модельного сигнала выполняется в следующей последовательности:

– строится график модельного сигнала;

– обнаруживаются пики;

– фрагменты данных, содержащие пики, удаляются;

– выполняется аппроксимация ориентировочной базовой линии монотонной функцией по данным, не содержащим пики;

– аппроксимированная базовая линия сравнивается с базовой линией модельного сигнала, оценивается погрешность построения базовой линии;

– аппроксимированная базовая линия вычитается из модельного сигнала для его последующей обработки.

Использованы следующие команды `MATLAB` [5]:

– `findpeaks()` — определение положения пиков (максимумов функции на интервале);

– `sgolayfilt()` — применение фильтра Савицкого—Голея;

– `polyfit()` — аппроксимация функции по методу наименьших квадратов полиномом степени n;

– `polyval()` — вычисление значения этого полинома;

– `interp1()` — построение интерполирующей кривой.

Оцененная абсолютная погрешность построения базовой линии не превосходит величину шума. Такой результат обеспечивает достижение минимальной погрешности определения положения пиков.

ФИЛЬТРАЦИЯ ШУМА

В генетическом анализаторе НАНОФОР-05 был применен способ нахождения вершин пиков, описанный в статьях [6, 7]. Фильтрация шума выполнялась сочетанием фильтров на основе медианной фильтрации и скользящего среднего значения (прямоугольное окно) по 3 точкам.

В данной статье выполнено сравнение эффективности таких фильтров и фильтра Савицкого—Голея путем вычисления среднеквадратического отклонения S (СКО). В верхней строке таблицы показано, что при максимальной амплитуде двуполярного шума, равной ± 100 о.е., без фильтра S = 57.70 о.е., а после фильтрации СКО уменьшается $S_f \leq S$.

Показано, что фильтр Савицкого—Голея значительно (более чем в 1.7 раз) повышает отношение сигнала к шуму при незначительном изменении высоты и ширины пиков.

Еще одним известным способом фильтрации является спектральная фильтрация, которая рассмотрена в следующем разделе данной статьи.

СКО (S, о.е.) шумового сигнала до и после фильтрации

Использование преобразования Фурье	Исходный шум	Фильтр		
		Медиана	Среднее по 3 точкам	Савицкого—Голея
Без преобразования Фурье	57.70	44.52	33.02	33.00
Двойное преобразование Фурье после фильтра	16.79	20.71	16.17	16.66
Фильтр после двойного преобразования Фурье	–	16.68	16.18	16.67

ВОЗМОЖНОСТИ ПОВЫШЕНИЯ РАЗРЕШАЮЩЕЙ СПОСОБНОСТИ

Предельное разрешение генетического анализатора (700 и более нуклеотидов) определяется возможностью разделения соседних пиков, принадлежащих фрагментам максимальной длины и отличающихся на 1 нуклеотид. Пики фрагментов максимальной длины значительно уширяются в основном за счет неравномерности температуры внутри капилляра в радиальном направлении и диффузии (время выхода последних пиков — порядка 1.5 ч).

Поскольку все молекулы фрагментов ДНК одинаковой длины, содержащиеся в исходной пробе, имеют одинаковые свойства, то форма пиков определяется так называемой аппаратной функцией. В первом приближении форму таких пиков можно аппроксимировать функцией Гаусса. После преобразования Фурье получается спектр, форма огибающей которого имеет вид спадающей экспоненты.

Известен способ частичного восстановления свойств исходной пробы (обострение пиков) путем учета аппаратной функции — способ деконволюции [8]. Для реализации этого способа спектр Фурье нужно разделить на спектр аппаратной функции. Повышения разрешающей способности можно достичь, если спектр Фурье сигналов фрагментов ДНК умножить на нарастающую экспоненту. После такой обработки ширина пиков уменьшается, а соседние пики частично разделяются. Недостатком такой простой реализации является значительное возрастание высокочастотных составляющих спектра, что вызывает увеличение шума.

Предлагается оптимизировать способ повышения разрешения пиков путем объединения его со способом фильтрации шума. Такой объединенный способ можно рассматривать как способ спектральной фильтрации.

В качестве одного из вариантов предлагается спектр Фурье умножить на корректирующую функцию в виде двух пиков, форма которых определяется функцией Гаусса. Оптимизация корректирующей функции достигается путем выбора двух параметров: координат максимумов кривой на спектральной оси и полуширины функции на уровне 0.6 (σ).

При выборе первого параметра в качестве ориентира можно использовать значение базового интервала $T_б = t_N – t_{(N–1)}$ между соседними пиками N и (N–1) фрагментов, отличающихся на 1 нуклеотид. При общей длине сигналов зависимости интенсивности флуоресценции от времени M положение гармоники $F_б$ на левой части графика спектра Фурье можно выразить формулой

$$F_б = M / T_б.$$

Величины M и $T_б$ в этой формуле выражены в виде количества точек исходного графика, величина $F_б$ соответствует номеру точки на графике спектра Фурье. Эта формула следует из определения преобразования Фурье: если за единицу времени принять интервал времени M, в течение которого брались входные данные, то разложение сигналов с периодом $T_б$ на синусоидальные составляющие соответствует частоте $F_б$. Например, если M = 3000 точек и $T_б$ = 20 точек, то $F_б$ соответствует 150 точке от начала графика спектра Фурье.

Ширина корректирующей функции подбирается, исходя из допустимого искажения формы пиков, в виде так называемых "крыльев". Далее будет показано, что одновременно автоматически достигается коррекция базовой линии, а ширина корректирующей функции влияет на степень такой коррекции.

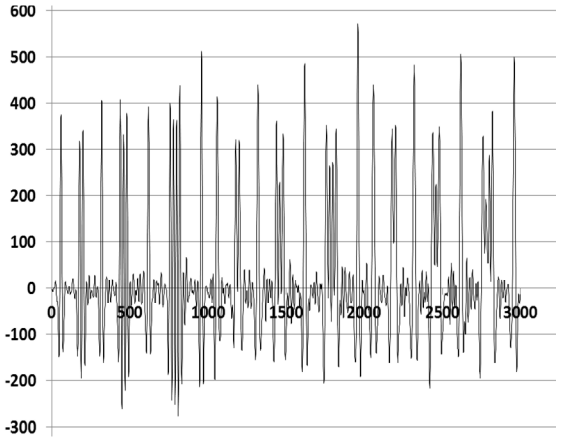

Рис. 2. Модельный сигнал после обработки. Оцифровка осей аналогична рис. 1

Для примененной модели сигналов фрагментов ДНК корректирующая функция имеет следующий вид:

```
B1=2.72^(-((X1-A1)^2)/(2*σ^2))+
+2.72^(-((X2-A1)^2)/(2*σ^2)),
```

где

2.72 — приближенное значение e;

`X1` = 150 — положение центра первой функции Гаусса;

`X2` = 2850 — положение центра второй функции Гаусса;

$\sigma = 75$ (для данного примера).

Модель корректирующей функции сохранялась в текстовом редакторе с именем файла x.txt.

Такой объединенный способ повышения разрешения пиков путем объединения его со способом фильтрации шума можно рассматривать как способ спектральной фильтрации, при котором преимущественно выделяется гармоника сигнала, соответствующая частоте $F_б$, которая приблизительно известна из предыдущих анализов.

При реализации предложенного способа в программе `MATLAB` были использованы следующие команды:

– `fo=fft(o)` — прямое преобразование Фурье любого модельного сигнала с именем "`o`";

– `ox=fo.*x` — умножение спектра Фурье модельного сигнала на корректирующую функцию x;

– `iox=ifft(ox)` — обратное преобразование Фурье одного из модельных сигналов после спектральной фильтрации;

– `riox =real(iox)` — реальная часть обратного преобразования Фурье.

В результате обработки получены следующие положительные эффекты:

– уменьшение ширины пиков на уровне 0.5 до 30 %;

– уменьшение шума — примерно в 3 раза;

– уменьшение дрейфа базовой линии — в 7 раз.

Отрицательными эффектами можно считать уменьшение сигналов (неравномерное, от 2 до 5 раз) и наличие остаточного узкополосного шума.

Для преодоления этих отрицательных эффектов предлагается применить следующие дополнительные операции:

– выполнить нормализацию амплитуд графика после спектральной фильтрации;

– для уменьшения влияния шума применить дополнительную фильтрацию (средняя и последняя строки в таблице);

– для декоративного улучшения вида графика после коррекция базовой линии или после Фурье выделить пики путем установки базовой линии выше уровня шума (обрезать шумы).

При оценке результата сочетания фильтров в средней строке таблицы обнаружен неожиданный эффект: использование медианной фильтрации перед спектральной фильтрацией шумового сигнала увеличивает стандартное отклонение (20.71 о.е.). Объяснить такой эффект можно за счет нелинейного преобразования шумовых сигналов и переноса высокочастотных шумов в область спектрального фильтра. Поэтому такое сочетание фильтров использовать нецелесообразно.

График модельного сигнала, объединяющий пики и шумовой сигнал, после спектральной обработки и нормализации амплитуд приведен на рис. 2.

ЗАКЛЮЧЕНИЕ

1. На основе предложенной модели сигналов фрагментов ДНК генетического анализатора выполнено сравнение известных способов коррекции базовой линии и фильтрации шума.

2. Предложен оптимизированный способ повышения разрешения пиков путем использования аппаратной функции, параметры которой определяются с учетом известных спектральных свойств сигнала генетического анализатора.

3. Показана возможность достижения одновременного повышения разрешения пиков, обеспечения коррекции базовой линии и фильтрации шума. Рассмотрены возможности сочетания разных способов обработки генетической информации.

СПИСОК ЛИТЕРАТУРЫ

1. ИАП РАН. Каталог приборов. Генетический анализатор НАНОФОР® 05. URL: (http://www.iai.rssi.ru/nanofor05.php).
2. *Белов Ю.В., Леонтьев И.А., Панчук В.В. и др.* Построение калибровочной линии при фрагментном анализе ДНК // Научное приборостроение. 2013. Т. 23, № 3. С. 26–31.
3. *Алексеев Я.И., Белов Ю.В., Малюченко О.П. и др.* Генетический анализатор для фрагментного анализа ДНК // Научное приборостроение. 2012. Т. 22, № 4. С. 86–92.
4. *Белов Ю.В., Леонтьев И.А., Петров А.И., Курочкин В.Е.* Коррекция базовой линии сигналов флуоресцентного детектора генетического анализатора // Научное приборостроение. 2013. Т. 23, № 2. С. 9–13.
5. *Потемкин В.Г.* Справочник по MATLAB. Анализ и обработка данных. URL: (http://matlab.exponenta.ru/ ml/book2/chapter8/ diff.php).
6. *Леонтьев И.А.* Обработка данных в задачах электрофореза // Научное приборостроение. 2003. Т. 13, № 2. С. 96–99.
7. *Леонтьев И.А.* Обсчет пиков в задачах электрофореза // Научное приборостроение. 2004. Т. 14, № 1. С. 94–96.
8. *Василенко Г.И.* Теория восстановления сигналов. М.: Советское радио, 1979. 272 с.

Институт аналитического приборостроения РАН, г. Санкт-Петербург

Контакты: *Манойлов Владимир Владимирович*, manoilov_vv@mail.ru;
Белов Юрий Васильевич, bel3838@mail.ru

Материал поступил в редакцию: 10.07.2014

UDK 543.426; 543.9

METHODS OF GENETIC ANALYSIS RESULTS PROCESSING

D. A. Belov, Yu. V. Belov, V. V. Manoylov, V. E. Kurochkin

Institute for Analytical Instrumentation of RAS, Saint-Petersburg, RF

The genetic signals simulation for assessment of digital signal processing efficiency was carried out. The possibilities of different methods of baseline correction, filtering and increasing resolution in genetic analyses were considered.

Keywords: DNA, genetic analyzer, fluorescent detection

REFERENCES

1. URL: (http://www.iai.rssi.ru/en/catalog.php).
2. URL: (http://matlab.exponenta.ru/ ml/book2/ chapter8/ diff.php).

Contacts: *Manoylov Vladimir Vladimirovich*, manoilov_vv@mail.ru;
Belov Yuriy Vasil'evich, bel3838@mail.ru

Article arrived in edition: 10.07.2014